UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Exploring Biomedical Ontologies, Personalized PageRank and Semantic Similarity in the Entity Linking task

Pedro Simões Ruas

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof. Doutor Francisco José Moreira Couto

2019

Resumo

A literatura científica está maioritariamente disponível na forma de artigos publicados, que são essenciais para a partilha de conhecimento científico. Contudo, o ritmo de publicação de novos artigos tem aumentado constantemente, excedendo a capacidade humana de gerir e aceder a esta grande quantidade de texto não estruturado: os investigadores despendem mais esforço e tempo a recuperar informação científica e o trabalho dos biocuradores torna-se mais complicado, pois a maior parte do texto não é estruturada, o que complica a aplicação de ferramentas automáticas. Os métodos de Prospeção de Texto podem ajudar a gerir a situação, mais concretamente, através da extração automática de informação a partir do texto presente na literatura científica. A tarefa de Mapeamento de Entidades, responsável por fazer corresponder entidades identificadas no texto a um conceito de uma Base do Conhecimento, é um passo essencial de muitos sistemas de Prospeção de Texto. Mas, comparando com outros domínios, como por exemplo o texto proveniente de redes sociais, a disponibilidade de ferramentas capazes de efectuar Mapeamento de Entidades é ainda escassa.

Esta dissertação propõe um módulo capaz de efectuar Mapeamento de Entidades em documentos anotados com entidades pertences a duas ontologias biomédicas: *Gene Ontology* (GO) e *Uber-Anatomy Ontology* (Uberon). O sistema utiliza o algoritmo PageRank personalizado e medidas de semelhança semântica para escolher o melhor candidato para cada entidade do texto. O desempenho do sistema foi avaliado no corpus CRAFT, alcançando uma eficácia de 0.8244 em entidades pertencentes à subontologia GO *Biological Process*, de 0.7258 em entidades da subontologia GO *Cellular Component* e de 0.7918 em entidades da ontologia Uberon. Adicionalmente, o sistema foi avaliado no corpus MSNBC que contém entidades da ontologia *DBpedia* e

alcançou uma eficácia de 0.8814, o que é comparável com resultados alcançados por sistemas estado da arte. O código do módulo pode ser consultado na página GitHub do grupo LaSIGE Biomedical Text Mining Team: `https://github.com/lasigeBioTM/PPRSSM`.

Os resultados do presente trabalho demonstram que é possível aplicar com sucesso medidas de semelhança semântica a sistemas baseados no algoritmo PageRank personalizado e explorar ontologias biomédicas para melhorar a tarefa de Mapeamento de Entidades.

Palavras Chave: Prospecção de Texto, Mapeamento de Entidades, Semelhança Semântica, PageRank Personalizado, Ontologias Biomédicas

Abstract

Scientific literature is mainly available in the form of published articles, which are essential to the sharing of scientific knowledge between researchers. However, the rate of publication of new articles have been steadily rising, exceeding the human capacity to effectively manage and assess this large amount of unstructured text: researchers spend more time and effort to retrieve scientific information and the task of biocurators also gets more difficult, due to the unstructured nature of the text that hinders the application of automatic tools. Text Mining methods can help to manage this situation, more concretely by automatically extracting information from the text in scientific literature. Entity Linking, the task of automatically mapping entities recognized in text to a knowledge base concept, is an essential step in Text Mining workflows. But, comparing to other domains like social media text, the availability of EL tools capable of performing well in biomedical text is still scarce.

This dissertation proposes a module that performs Entity Linking in documents annotated with entities from two biomedical ontologies: *Gene Ontology* (GO) and *Uber-Anatomy Ontology* (Uberon). The system applies the Personalized PageRank (PPR) algorithm and semantic similarity measures to choose the best candidate for each entity in text. The performance of the system was evaluated on CRAFT corpus (gold standard), achieving an accuracy of 0.8244 in GO *Biological Process* entities, 0.7258 in GO *Cellular Component* entities and 0.7918 in Uberon entities. Additionally, the system was evaluated on the MSNBC gold standard containing *DBpedia* entities and achieved an accuracy of 0.8814, which compares well with other state-of-the-art systems. The code behind the module can be accessed in the LaSIGE Biomedical Text Mining Team GitHub page: `https://github.com/lasigeBioTM/PPRSSM`.

The results of the present work prove that it is possible to successfully apply semantic similarity measures in PPR-based systems and explore biomedical ontologies for the improvement of the EL task.

Keywords: Text Mining, Entity Linking, Semantic Similarity, Personalized PageRank, Biomedical Ontologies

Resumo Alargado

A literatura científica é um dos meios privilegiados de comunicação em Ciência, pois grande parte dos investigadores continua a partilhar os resultados obtidos nas suas experiências sob a forma de artigos. A publicação de artigos possibilita que investigadores numa determinada área científica estejam a par do conhecimento que vai sendo produzido na sua área por investigadores pertencentes a outros grupos e a outras instituições. No entanto, o ritmo de publicação de novos artigos tem vindo a aumentar constantemente, pelo que se torna cada vez mais difícil acompanhar o que vai sendo feito nas diferentes áreas das ciências biomédicas e das ciências da vida. Muitos investigadores podem estar a despender tempo, esforço e dinheiro em experiências que já foram realizadas por outros colegas. O avolumar de publicações nos repositórios científicos dificulta a recuperação e o acesso à informação científica, especialmente se se considerar que grande parte do texto presente nestes documentos não se encontra numa forma estruturada. Isto faz com que ferramentas automáticas que poderiam ajudar nesses processos não possam ser directamente aplicadas sobre o texto. Por outro lado, o trabalho dos biocuradores de extrair informação a partir de artigos para armazenamento em repositórios de informação biológica torna-se gradualmente mais complexo e moroso, pois há mais texto que necessita de ser analisado.

Os métodos de Prospeção de Texto visam a extração automática de informação a partir de grandes quantidades de texto não estruturado, pelo que podem fazer parte da solução para o problema da acumulação de publicações científicas. As ferramentas que efectuam Mapeamento de Entidades são cruciais no contexto dos sistemas de Prospeção de Texto. Mapeamento (ou desambiguação) de Entidades é uma tarefa que tem por objectivo fazer corresponder entidades identificadas num

dado texto a conceitos de uma base de conhecimento (como uma ontologia). A área das ciências biomédicas e das ciências da vida carece de ferramentas capazes de efectuar Mapeamento de Entidades, especialmente se se comparar com outras áreas mais exploradas, como é o caso do texto proveniente de redes sociais. Por este motivo, o presente trabalho considera que o desenvolvimento de ferramentas que efectuem Mapeamento de Entidades em documentos científicos é absolutamente essencial e que o trabalho dos investigadores científicos e dos biocuradores pode ser beneficiado com a adoção deste tipo de ferramentas.

O algoritmo PageRank personalizado (PPR) tem aplicações documentadas na tarefa de Mapeamento de Entidades, sobretudo ao nível da desambiguação de entidades pertences à *Wikipedia*. Este é um método baseado em grafos que considera a tarefa de Mapeamento de Entidades como uma tarefa de classificação (*ranking*), em que para cada entidade ou menção textual é construída uma lista de possíveis candidatos a partir da base de conhecimento considerada. Com os candidatos para todas as entidades é então construído um grafo, em que cada cada candidato constitui um nó e as ligações entre nós são adicionadas de acordo com as relações existentes entre os respectivos candidatos no contexto de uma base de conhecimento. O candidato que mais contribui para maximizar a coerência global do grafo, i.e., o candidato que se ajusta melhor no contexto do grafo, irá ser selecionado como a desambiguação correta para a respectiva entidade.

Uma ontologia é uma representação estruturada de uma parte da realidade que normalmente inclui uma lista de conceitos, as suas definições e as relações que existem entre os conceitos. As ontologias biomédicas representam, como o seu próprio nome sugere, uma parte do conhecimento biomédico. O exemplo mais conhecido é a *Gene Ontology*, que inclui conceitos relacionados com funções de genes e proteínas e apresenta uma estrutura de grafo acíclico direcionado, ou seja, as ligações entre nós têm sempre uma direção e um determinado nó nunca tem uma Mapeamento para si próprio. Outro exemplo

de ontologia biomédica é a *Uber-Anatomy Ontology* que se dedica à representação de conceitos relacionados com partes anatómicas. As relações no contexto desta ontologia são de subsunção, por exemplo, uma subclasse B está incluída numa superclasse A e assim sucessivamente. A estrutura destas ontologias pode ser explorada na tarefa de Mapeamento de Entidades, nomeadamente, na construção das ligações do grafo de candidatos e na criação de uma lista de candidatos para cada entidade presente num texto.

As medidas de semelhança semântica (SSMs) determinam o grau de informação partilhada entre dois conceitos de uma ontologia. Podem por isso ser usadas na determinação da semelhança semântica entre dois nós no grafo de candidatos, introduzindo assim o conceito de coerência local. Assim, a classificação de um nó/candidato no grafo passa a ser baseada na sua contribuição para a coerência global e na sua coerência local com os outros nós do grafo.

Deste modo, o principal objectivo da dissertação foi construir um módulo capaz de efectuar Mapeamento de Entidades em artigos científicos na área das ciências biomédicas, explorando para o efeito o algoritmo PageRank personalizado, medidas de semelhança semântica e a estrutura de duas ontologias biomédicas, *Gene Ontology* (GO) e *Uber-Anatomoy Ontology* (Uberon). O módulo foi integrado no sistema PPR-SSM, desenvolvido por André Lamúrias, que efectua Mapeamento de Entidades pertencentes a duas ontologias biomédicas: *Chemical Entities of Biological Interest Ontology* (ChEBI) e *Human Phenotype Ontology*. O código desenvolvido para o módulo encontra-se disponivel na página GitHub do grupo LaSIGE Biomedical Text Mining Team: `https://github.com/lasigeBioTM/PPRSSM`.

O desempenho do módulo foi avaliado no corpus "Colorado Richly Annotated Full-Text" (CRAFT), que é constituído por artigos científicos anotados com entidades pertencentes, entre outras, à GO e à ontologia Uberon. Para além disso, como a maior parte das ferramentas de Mapeamento de Entidades existentes se dedicam à desambiguação de entidades da *Wikipeda*, o desempenho do módulo também

foi avaliado no MSNBC corpus, uma coleção de vinte artigos de notícias. Neste último, a desambiguação de entidades foi efectuada para a ontologia *DBpedia*, que é a versão estrurada da *Wikipedia*. Assim, a comparação com o desempenho de outros sistemas estado da arte torna-se possível.

Relativamente ao corpus CRAFT, este foi separado em três conjuntos distintos: CRAFT-BP (anotações pertencentes à subontologia *Biological Process* da GO), CRAFT-CC (anotações pertencentes à subontologia *Cellular Component* da GO) e CRAFT-UB (anotações pertencentes à ontologia Uberon). Para cada um destes conjuntos, foi calculada a eficácia da desambiguação, que corresponde ao número de entidades correctamente desambiguadas a dividir pelo número total de entidades desambiguadas.

O módulo alcançou os seguintes valores de eficácia: 0.8244 CRAFT-BP, de 0.7258 no CRAFT-CC e de 0.7918 no CRAFT-UB. Estes valores constituem um aumento de eficácia comparativamente aos modelos base usados: um modelo base que consistia em escolher o candidato morfologicamente mais semelhante à entidade considerada, um outro modelo que apenas aplicava o algoritmo PPR sem SSMs e um terceiro modelo que aplicava o algoritmo PPR com o conceito de *Information content*, que mede o grau de "raridade" de um determinado conceito num contexto de um corpus.

O módulo alcançou uma eficácia de 0.8814 no corpus MSNBC, maior do que aquela alcançada pelos modelos base e que é comparável aos valores obtidos por outros sistemas estado da arte. Este resultado demonstra que o módulo desenvolvido pode ser aplicado em áreas para além do domínio biomédico.

Os resultados do presente trabalho demonstram que é possível explorar SSMs e ontologias biomédicas para melhorar a eficácia dos métodos de Mapeamento de Entidades baseados no algoritmo PPR.

Agradecimentos

Ao professor Francisco Couto, pela oportunidade que me deu de trabalhar no seu grupo e por todas as ideias, críticas construtivas e conselhos que foi providenciando ao longo da elaboração da dissertação.

Ao André Lamúrias, cujo trabalho previamente realizado foi o ponto de partida para a elaboração desta dissertação. Para além disso, esteve sempre disponível para discutir ideias e fazer sugestões que contribuíram significativamente para a realização desta dissertação.

Ao LaSIGE e à FCT, pelo apoio financeiro prestado. Este trabalho foi financiado pela FCT através do projecto "DeST: Deep Semantic Tagger", ref. PTDC/CCI-BIO/28685/2017 (`http://dest.rd.ciencias.ulisboa.pt/` e pela unidade de investigação LASIGE, ref. UID/CEC/00408/2019

Aos meus pais, aos meus irmãos, aos meus avós e ao Arménio por me apoiarem sempre a todos os níveis e por se mostrarem interessados no trabalho que aqui desenvolvi.

À Filipa, por estar presente todos os dias e por me apoiar incondicionalmente. Sem ela não teria concluído esta etapa da minha vida.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Scientific literature is the major source of scientific knowledge and it is available in the form of journal articles, patents, scientific reports, and in other formats. For instance, journal articles constitute the privileged medium to share the results obtained in scientific experiments and to advance new ideas and theories, as well to discuss the existing ones. In this way, published articles play a crucial role in the communication between researchers that share an interest about a certain scientific topic but are physically distant from each other. Additionally, if someone wants to assess the state-of-the-art in a scientific field, usually the best option is starting to analysing the already published articles (Couto, 2019).

However, the rate of publication of new articles exceeds the human capacity to read and to analyse them. Looking for MEDLINE statistics (accessible through PubMed)[1], it is possible to observe that by the end of 2018 fiscal year (30 September 2018) 25,239,968 citations were indexed in the repository, 904,636 alone corresponding to additions made in that year. For researchers, the task of keeping up-to-date on their scientific topic is becoming increasingly laborious, if not impossible, thus cancelling out the advantages provided by journal articles. On the other side, the work of a biocurator responsible for transferring the knowledge contained in scientific publications to a biological repository also gets more

[1] `https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html`

difficult. Most of the text in these articles is expressed in natural language, so it is unstructured by nature, which hinders the application of automatic methods over the text that could assist in the biocurator's task (Cohen & Hunter, 2007).

Text mining (TM) methods can help to put this messy landscape into order, more specifically by automatically extracting information from a large amount of unstructured text. According to Fleuren & Alkema (2015), a text mining workflow usually includes the following steps:

1. Information retrieval, with extraction of the relevant documents from multiple literature sources.

2. Named entity recognition and linking, including the extraction of entities in text and their linking to the appropriate concepts in a knowledge base (like an ontology).

3. Extraction of relationships between concepts in the text

4. Knowledge discovery, for example inferring new relationships between concepts based on existing knowledge or generating new scientific hypotheses.

Many works have been done in the field of biomedical and life sciences TM. For example, Singhal *et al.* (2016) proposed an automated machine learning system to extract gene-disease relations described in literature which has applications in both precision medicine and in database curation. More recently, Lamurias *et al.* (2019) developed BO-LSTM, a model that is capable of extracting relationships between concepts belonging to several ontologies, such as *Chemical Entities of Biological Interest* (ChEBI), *Human Phenotype Ontology* (HPO) *and Gene Ontology* (GO). Müller *et al.* (2018) proposed Textpresso Central, an information retrieval system that assists online literature search and biocuration. Kilicoglu (2018) argues that TM systems could be used in the promotion of rigor and integrity in biomedical research through detection of fraud or plagiarism and could help to overcome the lack of science reproducibility: being able to manage the huge amount of scientific literature available can decrease the waste of time and money on repeated scientific experiments that are already described in the literature. Approaches to named entity recognition have also been suggested

(Basaldella *et al.*, 2017; Munkhdalai *et al.*, 2015), however the number of systems performing entity linking (EL) or disambiguation still lags behind. The gap widens if we make a comparison with the number of systems devoted to EL in social media (Derczynski *et al.*, 2015; Shen *et al.*, 2013). EL in the biomedical and life sciences presents additional challenges comparing to other domains, as the text is full of abbreviations and acronyms that have several possible resolutions (Hunter & Cohen, 2006), usually requiring expert knowledge to deal with the ambiguity.

Personalized PageRank (PPR) is a variation of the algorithm *PageRank*, initially developed by Google in 1999. PPR has some documented applications to EL task, like Pershina *et al.* (2015), that consist in a PPR graph-based method to perform disambiguation of *Wikipedia* entities. The nodes of the graph correspond to candidates for the entities (or to *Wikipedia* pages) and the system leverages the hyperlink structure of *Wikipedia* to generate the edge structure of the graph. This approach considers EL as ranking task, in which PPR performs random walks in the graph and ranks each node according to its contribution to the global coherence of the graph. This means that higher ranking nodes will fit better in the graph than lower ranking ones, constituting probable candidates for the respective entities. Consequently, the highest ranking node for each entity is chosen as the correct disambiguation.

An ontology is a structured, graph-theoretic representation of the reality that contains a set of concepts, their definitions and the relationships between them (Gruber, 1993). Domain ontologies focus on a specific part of the reality, as it is the case of the biomedical ontologies. *Gene Ontology* (GO), for example, represents concepts associated with gene and protein functions. GO and other biomedical ontologies have gained an increasing attention in the last years, as they confer undeniable advantages in the management, standardization and sharing of scientific knowledge (Arp *et al.*, 2015). In this way, it is important that the scientific knowledge present in published articles can be effectively stored in ontologies to allow further exploration by automatic methods. Besides that, the vast knowledge stored in ontologies can be leveraged in the development of EL systems. For example, the structure of the candidates graph generated for application of PPR

algorithm can be improved according with the semantic relationships expressed in the ontology, because each candidate is an ontology concept.

Semantic similarity aims at calculate the degree of shared information between two given concepts (of an ontology, for example). Biomedical ontologies, like GO, provide structured representations of scientific knowledge which allows the application of semantic similarity measures (SSMs) in order to compare two biological entities (Couto & Lamurias, 2018b). The advantages of SSMs can be explored in the candidate-ranking step of EL systems.

1.2 Objectives

As EL plays a crucial role in TM workflows, this work considers that there is an urgent need of systems capable of effectively performing EL in biomedical and life sciences text and that both researchers and biocurators would see their work facilitated with the availability of these tools.

Considering the framework proposed by Pershina *et al.* (2015) that applies the PPR algorithm to the EL task, and the advantages documented for the use of biomedical ontologies and SSMs, the objective of the present work is to develop a module capable of performing EL in biomedical articles leveraging these resources. PPR application will be improved with the application of SSMs, as these will allow a better estimation of the coherence of each node in the candidates graph, thus reducing the overall amount of wrong disambiguations. If a given node of the graph is highly similar with many other nodes, which can be assessed through SSMs application, it will contribute more to the global coherence of the graph. The knowledge encoded in two biomedical ontologies, GO and Uber-anatomy Ontology (Uberon represents anatomical parts), will be used to generate a list of candidates for each entity recognized in a document and to build the edge structure of the candidates graph. This contrasts with the method proposed by Pershina *et al.* (2015), where the edge structure is generated after *Wikipedia* hyperlink structure and where SSMs are not used. The ultimate functionality of the proposed module will be to provide a distinct ontology identifier to each entity recognized in text.

Hypothesis: It is possible to improve the performance of entity linking in biomedical articles using Personalized PageRank, semantic similarity and the knowledge encoded in domain ontologies, like *Gene Ontology* and *Uber-anatomy Ontology*

The performance of the developed tool will be evaluated on the "The Colorado Richly Annotated Full-Text" (CRAFT) corpus, a dataset containing biomedical articles annotated with GO and Uberon entities. As many EL systems are primarily built for the disambiguation of *Wikipedia* entities or other entities belonging to general domains, the EL module here developed will additionally be tested on the MSNBC corpus, a collection of news stories annotated with *Wikipedia* entities, for comparison with state-of-the-art systems..

1.3 Contributions

The main contribution of this work is a biomedical EL tool:

Biomedical Entity Linking module (PPR-SSM): Development of a module capable of performing disambiguation of *Gene Ontology* and *Uber-anatomy Ontology* concepts in biomedical articles. Integration of the module in the PPR-SSM system built by André Lamúrias. The software is available at the LaSIGE Biomedical Text Mining Team GitHub page:

`https://github.com/lasigeBioTM/PPRSSM`

The development and evaluation of PPR-SSM method in several datasets originated a paper that was submitted to publication:

LAMÚRIAS, A., RUAS, P., COUTO, F.M.(2019). PPR-SSM: Personalized PageRank using Semantic Similarity Measures for Entity Linking

1.4 Overview

The overview of this document is as follows.

Chapter 2 explains the concepts that are needed to understand the theory behind this work, as well a brief overview of the existent EL systems.

Chapter 3 exposes the theoretic foundations for the methods used in this work, describing the different models tested.

Chapter 4 provides an overview of the data, the evaluation setup and the results achieved in the different datasets. Additionally, there is a discussion about the main errors associated with the performance of the system, as well about the implications of the achieved results.

Chapter 5 presents the main conclusions extracted from this work, as well some approaches to improve the proposed system in the future.

Chapter 2

Background

2.1 Entity linking

Entity linking (EL), which can also be designated by entity disambiguation or normalization, is a natural language processing task that links named entities present in text to the appropriate entry (or entries) in a knowledge base (KB) (Shen *et al.*, 2015). Usually, named entity recognition task precedes EL, identifying the named entities in the text and the respective boundaries. EL plays an important role in text-mining applications, including the linking of entities in social media text (like Twitter) to Wikipedia or other KB (Gattani *et al.*, 2013), the population of KB with entities extracted from text (Dredze *et al.*, 2010) and the linking of entities in web search queries (Blanco *et al.*, 2015).

The development of EL systems and its integration in broader text-mining pipelines have plenty applications in biomedical and life sciences domains:

- Automated document classification and document retrieval, for example, to help the researchers to find the most relevant articles to their research (Jovanovi & Bagheri, 2017)
- Automated biocuration, reducing the human effort and time spent in the process (Rak *et al.*, 2014)
- Data integration between different repositories (Perez-Riverol *et al.*, 2017)

- Computational modelling, for example it is possible to model a pathway comprising several biological entities that are linked to biological repositories (Zheng *et al.*, 2015) or to build networks around a biomedical entity (Lee *et al.*, 2016a).

Some of the EL systems are incorporated in specialized search engines with the aim of improving the quality and efficacy of queries results, as it is the case of the biomedical entity search tool *BEST* (Lee *et al.*, 2016b) and BioSearch (Hu *et al.*, 2017) and others are integrated in platforms that combine entity linking and other text mining features that perform in scientific literature, like *DeepLife* (Ernst *et al.*, 2016) or *BeCas* (Nunes *et al.*, 2013).

Another field that can benefit from EL application is clinical text enrichment. This type of text, present in electronic health records (EHR) for example, is largely expressed in an unstructured manner. In consequence, the analysis of the content is necessarily manual. Physicians spend two times more of their work hours managing EHR and doing other administrative tasks than directly assisting patients (Sinsky *et al.*, 2016). Through clinical text processing and EL, the normalized entities in the text can be used to search patient related information in repositories or in scientific literature or can be used in the summarization of clinical reports (Jovanovi & Bagheri, 2017). This allows the enrichment of patients data with information from different repositories while reducing the amount of time spent in information research and analysis. Ultimately, physicians will have more available time to contact directly with patients. Examples of works in this field are Kang *et al.* (2012), He *et al.* (2011) and Leaman *et al.* (2015).

According to Rao *et al.* (2013), some challenges associated with the development of EL systems are:

- **Entity name variations**, like abbreviations, acronyms, alternate spellings or synonyms. This is specially evident for gene nomenclature, where it is not uncommon for a single gene to have several symbols. Take for example the human gene `AFF1` which encodes the protein `AF4/FMR2 family member 1`[1]. Alternative symbols for this gene that appear in literature are `AF4`,

[1] `https://www.ncbi.nlm.nih.gov/gene/4299`

PBM1, MLLT2. The medical nomenclature is also prone to name variations. For example, the terms "myocardial infarction" and "heart attack" refer to the same medical condition.

- **Entity ambiguity**, because polysemous words can map to multiple KB concepts depending on the context. For example, the word *iris* can designate a circular structure of the eye in mammals and birds (anatomical organ) or can refer to a genus containing species of plants (taxonomic classification).

- **Absence of KB entry for a named entity**.

In biomedical text, these hurdles become harder to overcome since its resolution sometimes require domain expertise. Additionally, there is a lack of labelled data for EL in biomedical domain comparing to more explored domains, like news and social media, and the systems built for these domains do not perform well in biomedical text (Zheng *et al.*, 2015).

Some closely related tasks include:

- Word sense disambiguation identifies the meaning of a word in a given context making use of external knowledge sources, such as thesauri, machine-readable dictionaries and ontologies or corpora (Navigli, 2009).

- Co-reference resolution performs entity linking or disambiguation without a KB, creating clusters of entity mentions in one or more documents (Clark & Manning, 2016).

- Record linkage finds correspondence between records present in different databases, files or other sources to allow the integration of data in a coherent manner (Winkler, 1999).

2.2 Ontologies

The expanding use of computers in scientific disciplines, particularly in the realm of life sciences, has been massively increasing the amount of scientific data and

information available. This data influx derives from different techniques, from different research groups belonging to different fields or organizations in different parts of the world, which hampers the integration with the already existent information. Consequently, data access and data sharing becomes harder, as challenges related to storage, retrieval and reuse keep rising. Thus, it is necessary an unifying framework to ensure the "shared understanding" of scientific information. A given set of objects, concepts or other entities, as well the relationships between them, that exist in a given domain or in a part of the reality constitute a *conceptualization*, which is an abstract, schematic view of that domain or reality. An ontology is "an explicit specification of a conceptualization" (Gruber, 1993), representing in a formal manner a vocabulary of concepts, its definitions and the relationships between them. According to Arp *et al.* (2015), the ontologies are designed to ensure "consistency in description of data", in the sense that every concept has a textual definition, meaning that it is human readable and that consistence can be ensured by maintainers and users, and has a logical definition, which can be read and interpreted by automatic tools. The structure of an ontology is usually graph-theoretic, where the concepts are the nodes and relationships between concepts are the edges.

Domain ontologies represent a part of the knowledge in which the definition of the concepts is domain-specific, like in the case of biomedical ontologies. Their advantages are the common access to the data/information across domain and research groups boundaries, the integration of new scientific information with the one already existent, and the development of automatic mining tools which allow computer reasoning and can potentially unveil new scientific hypothesis (Arp *et al.*, 2015; Uschold & Gruninger, 1996).

The *Open Biological and Biomedical Ontology* (OBO) foundry is a collaborative initiative containing several inter-operable domain ontologies that are implemented according to a well defined set of principles and guidelines (Courtout *et al.*, 2011; Smith *et al.*, 2007). According to the proponents, the key requirements for integration in OBO are "that ontologies be open, orthogonal, instantiated in a well-specified syntax and designed to share a common space of identifiers" (Smith *et al.*, 2007). *Gene Ontology* (Section 2.2.1), *Chemicals of Biological Interest* on-

tology or ChEBI (Section 2.2.4) and *Uber-anatomy* ontology (Section 2.2.2) are part of the OBO foundry.

2.2.1 Gene Ontology (GO)

The *Gene Ontology Consortium*, originally composed by researchers studying the model organisms *Drosophila melanogaster* (fruit fly), *Saccharomyces cerevisiae* (yeast) and *Mus musculus* (mouse), created the GO[1] in 1998. The goal of this project was to unify the representation of gene and proteins roles in all eukaryotes (The Gene Ontology Consortium *et al.*, 2000), allowing, for example, the automatic transfer of biological annotations of more known organisms to less known ones. GO is a structured, hierarchical, controlled vocabulary for functions of genes and genes products in all organisms. It is a crucial bioinformatics resource that keeps evolving with the contributions of the community, including dedicated biocurators and experimental biologists (The Gene Ontology Consortium, 2019; The Gene Ontology Consortium *et al.*, 2000).

GO is a direct acyclic graph (DAG), where the nodes represent gene functions and edges represent relationships between the concepts. The relationships types represented in the ontology are: *is a*, *part of*, *has part*, *regulates*, *negatively regulates* and *positively regulates*. Figure 2.1 is a GO fragment representing GO terms and relationships between them.

GO represents 45,013 terms[2] distributed by three independent sub-ontologies: `Biological Process` (29,694 terms), `Molecular Function` (11,113 terms) and `Cellular Component` (4206 terms). An annotation is a connection between a gene product and one or more functions (terms) in GO. Each GO annotation has an evidence code that indicates the support for the annotation. There are several evidence codes, the main distinction laying down on manually validated annotations (`experimental evidence`, `phylogenetic evidence`, `computational evidence`, `author statements`, `curatorial statements`) versus `automatically generated annotations` (that are not manually validated)[3].

[1] `http://geneontology.org/`
[2] In February, 13, 2019
[3] `http://geneontology.org/docs/guide-go-evidence-codes/`

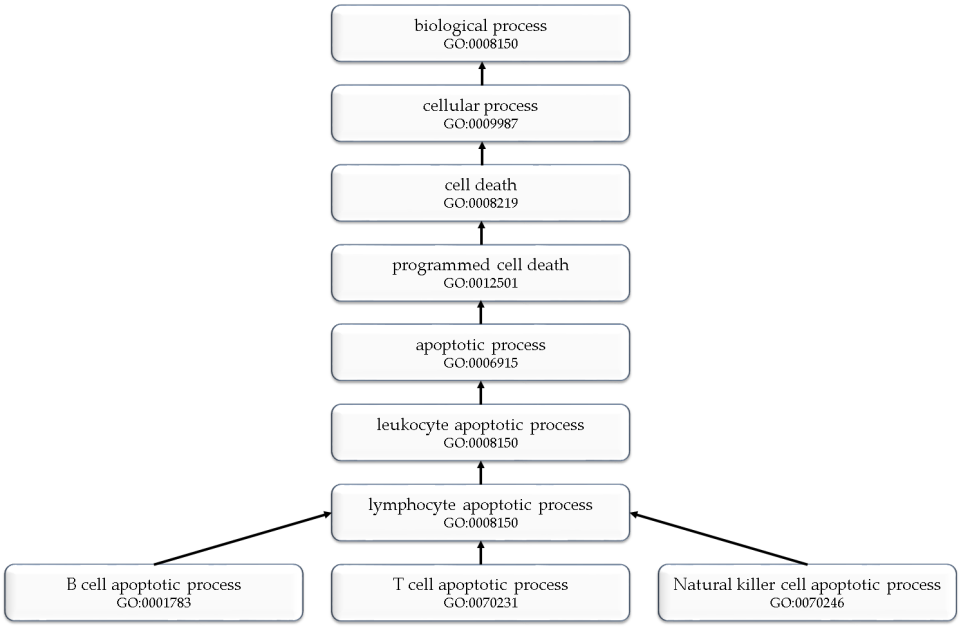


Figure 2.1: GO fragment with the terms GO:0001783, GO:0070231 and GO:0070246 and the respective ancestor terms up to the root term GO:0008150. Black arrows represent subsumption relationships

2.2.2 Uber-anatomy Ontology (Uberon)

Uberon [1] represents anatomical entities in a species-neutral way, as well the structural and developmental relationships between them (Mungall *et al.*, 2012). Currently, Uberon contains 13815 distinct classes [2].

The main goal of Uberon is to connect biological datasets annotated with different ontologies. There are two versions of the ontology: the main ontology, that contains exclusively Uberon concepts and the merged ontology, that contains relationships with concepts belonging to other ontologies, like GO, ChEBI, *Neuro Behaviour Ontology*, *Cell Ontology* and *Protein Ontology*. The main ontology has relationships of the type *is-a*, *part-of* and *develops-from*, whereas the merged ontology contains additional relationships, like *capable-of*, *has-function-in*, *has-part*.

Uberon ontology acts like a bridge because integrates data isolated in many

[1] `http://uberon.github.io/`
[2] In April, 14, 2019

species-centric anatomy ontologies (Haendel *et al.*, 2014), which allows, for example, automatic inference between entities of different taxa.

2.2.3 DBpedia

Wikipedia is a comprehensive encyclopedia about a variety of topics, including articles related with life sciences. For example, *Wikipedia* pages are widely accessed for its medical content either by the medical community or the general public and, in fact, this resource contained 30,000 articles on medical topics (considering the english version) as of March 2017 (Shafee *et al.*, 2017). Besides this, there are more examples of the *Wikipedia*'s scientific relevance, such as the *Gene Wiki* initiative, which is an open-source collection of 10,000 wiki articles about human genes (Good *et al.*, 2012) and the *RNA WikiProject*, an effort that created articles pertaining to 600 families of non-coding RNAs (Daub *et al.*, 2008).

The *DBpedia* project[1] is a community-based initiative with the aim of extract structured information from various editions of *Wikipedia*. The english version of the *DBpedia* ontology has 4,233,000 million instances mapped in 685 classes (3.7 release[2]).

2.2.4 Chemical Entities of Biological Interest Ontology (ChEBI)

ChEBI[3] is a database and ontology for representation of low-molecular weight chemical entities that are in some way associated with biological processes in living organisms (Hastings *et al.*, 2016). More specifically, ChEBI represents molecular entities (atoms, molecules, ion, ion pair, radical, radical ion, complex, etc.) that are either biochemical compounds or synthetic products, such as agrochemicals, laboratory reagents and pharmaceuticals. ChEBI ontology contains relationships of type *is-a*, *has-role*, *has-part*, among others, including 55,660 distinct entries[4]. ChEBI is integrated with GO, since many processes represented

[1] `https://wiki.dbpedia.org/`
[2] `https://wiki.dbpedia.org/services-resources/ontology`
[3] `https://www.ebi.ac.uk/chebi/`
[4] `https://www.ebi.ac.uk/chebi/statisticsForward.do`, in 1, April, 2019

by GO terms include the respective ChEBI entities that are involved in them (Hastings *et al.*, 2013).

2.2.5 Human Phenotype Ontology (HPO)

HPO[1] is a controlled vocabulary of phenotypical abnormalities associated with human diseases. It includes an ontology for phenotypes that are relevant in the medical field and disease-phenotype annotations. The ontology categorizes 13,000 terms in a DAG connected by *is-a* edges. Each one of the terms is assigned to one of the five sub-ontologies: `Phenotypic abnormality`, `Mode of Inheritance`, `Clinical modifier`, `Clinical course` or `Frequency`. HPO has relevant applications in personalized medicine, namely in computational deep phenotyping (computational analysis of details about disease manifestations at individual level and integration with the information and data available from other sources) and phenotype-driven genomic diagnostics (Sebastian *et al.*, 2019).

2.3 Semantic similarity

In a broad sense, semantics is the meaning of a word and semantic similarity is the shared meaning between two words. However, in biomedical domain, a word is usually a biological entity (gene, protein, compound, etc.) and its semantics is the biological function that it performs in a given biological context (Couto & Lamurias, 2018b)

How can we assess the semantic similarity between two biological entities? The first thing to keep in mind is the complexity and ambiguity associated with this task, because a biological entity can have a different function (or meaning) depending on context. Second, similar structure not always equate with similar function in biology. One paradigmatic example of this is the impact of Single Nucleotide Polymorphisms (SNP), where the modification of a mere "letter" (nucleotide) in "thousands of letters" (genomic sequence) can dictate the difference between being healthy or develop a genetic disorder and a consequent medical condition (Shastry, 2009).

[1] `https://hpo.jax.org/app/`

How can we measure semantic similarity? In opposition to humans, computers cannot assess semantic similarity in free text due to the very ambiguous nature of it. Thus, a semantic base is necessary to provide an unambiguous context, such as hierarchized common vocabularies, ontologies, taxonomies, or any other structured representation provided that it expresses semantic relationships between concepts.

Through semantic similarity measures (SSMs) it is possible to determine the degree of shared meaning or information in common between two concepts in a semantic base. SSMs usually only apply to subsumption (*is-a*) relationships in the semantic base (`cellular membrane` is a `cellular component`, `subclass B` is a `superclass A`, etc), even if there are present other types of relationships.

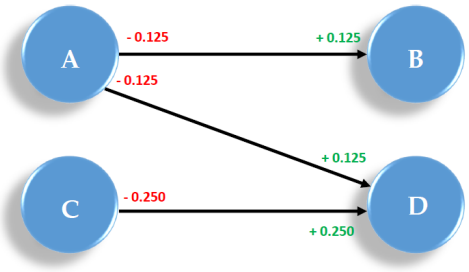
In biomedical sciences domain, semantic similarity has documented applications, for example, in information retrieval. Hliaoutakis *et al.* (2006) proposed a method capable of detecting the degree of similarity between two distinct documents according to the presence of MeSH (Medical Subject Heading) terms[1], even if the documents do not share lexically similar terms. In the same line, Alonso & Contreras (2016) presented a system to perform information retrieval in electronic health records, leveraging the Unified Medical Language System (UMLS) Metathesaurus[2] and SSMs.

2.4 PageRank

PageRank is a computational method to measure the relative importance of pages in World Wide Web. Initially proposed in 1999 (Page *et al.*, 1999), it aimed to improve information retrieval in the Web, in a scenario with increasing number of web pages and information heterogeneity. It was the basis of *Google*'s search engine.

Considering the web as a graph, each page (or node) has a variable number of forward links or out-edges pointing to other pages and a number of backlinks or in-edges coming from other pages. The PageRank algorithm returns a probability distribution of reaching web pages after successive iterations.

[1] `https://www.nlm.nih.gov/mesh/meshhome.html`
[2] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/`



	Iteration 0	Iteration 1
Page A	0.250	0.000
Page B	0.250	0.125
Page C	0.250	0.000
Page D	0.250	0.375

Figure 2.2: Simplified PageRank calculation for an example with four web pages. Left: schema with four web pages and respective link structure. Values in red represent PageRank decrease and values in green represent PageRank increase after the first iteration. Right: PageRank scores for each page at iteration 0 and iteration 1

Let's consider a simple network composed of four web pages **A**, **B**, **C** and **D** with the link structured depicted in Figure 2.2.

Initially the probability of reaching any given page or the PageRank score in this example is 0.250. At each iteration, a given page will transfer a PageRank score to each of its target page (pages to where the out-edges point to) and this score is equivalent to the initial value of the page divided by the number of out-edges. In Figure 2.2, page A has a score equal to 0.250 and two out-edges, so it transfers 0.125 to each of its target pages.

The PageRank is a "random surfer model" (Page *et al.*, 1999) based on random walks on a graph. The random surfer clicks randomly on successive links and, over a sufficiently large time, he finishes his trip in a given page. The PageRank score of a page represents the likelihood of terminate the trip in that page.

Sometimes, if the surfer ends in a small loop of web pages (pages that only point to each other) or in sink pages (pages without no forward links, like the pages B and D in Figure 2.2) eventually will get bored and thus will jump to a random web page without following a forward link. This behavior is modelled through the damping factor. For a damping factor of 0.85, the random surfer has an 0.85 likelihood of choosing a random forward link of the current page and a 0.15 likelihood of jumping to a random page in the graph. In Personalized PageRank (PPR), this jump always happens to a chosen page or node, hence the "personalized" in the algorithm denomination (Agirre & Soroa, 2009). The EL

task can be modelled as a ranking task, where for each entity mention in a document there is list with potential KB candidates (the nodes in a graph) and links between candidates (the graph edges). The candidates/nodes are then ranked according to the likelihood of being the correct disambiguation for the entity. There are documented adaptions of PPR to EL, such as Agirre & Soroa (2009), Pershina *et al.* (2015), Mazaitis *et al.* (2014). Details about PPR adaptation to EL are addressed in Chapter 3.

In a given document, modelling EL as a ranking task originates a graph with candidates for each entity as nodes. Each node represents a KB concept and links between nodes are added to the graph according to the relationships between the concepts in the KB. The system proposed in the present work explores semantic similarity in the context of ontologies to improve the node ranking in the graph jointly PPR (see Chapter 3).

2.5 State of the art

The challenges directed to researchers that apply text mining to biological problems are an effective way of assessing the state of the art in EL.

The BioCreAtIvE ("Critical Assessment of Information Extraction in Biology") challenge is one example. Its aim is to assess the state of the art in a set of tasks that have biological relevance and are related with biomedical digital curation (Hirschman *et al.*, 2005). The latest edition, BioCreAtIvE VI, was held in 2017 and comprised the "Interactive Bio-ID Assignment (IAT-ID) Track on innovations in Biomedical Digital Curation" that included the "Bioentity normalization task" [1]. The bioentities in the challenge dataset belonged to *Cellular component* GO subontology, ChEBI, Uberon, Cell Ontology, among others and the results of the participating systems in this challenge are available in Arighi *et al.* (2017).

One of the main topics of EL in biomedical domain is gene and protein name normalization. This stems from the fact that a gene or a protein can have multiple designations, for example, the human gene `PhospholipaseA2 groupVII`

[1] `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-1/`

can be designated by PLA2G7, PAFAD, PAFAH, LP-PLA2 or LDL-PLA2 [1]. Different researchers use different terminology in their publications which hampers both human reader and computer comprehension. BioCreAtIvE has organized tasks for the extraction of gene and protein mentions and respective linking to KB since its first edition in 2004 (Hirschman *et al.*, 2005), as well in other editions (Arighi *et al.*, 2017; Lu *et al.*, 2011; Morgan *et al.*, 2008). Other systems have been proposed, like Sullivan *et al.* (2011), Huang *et al.* (2011), Li *et al.* (2013) or Wei *et al.* (2015).

Another community challenge is the "BioNLP Shared Task series" (BioNLP-ST), whose last edition ocurred in 2016. The "Bacteria biotope (BB3) task"[2] required as first step the extraction of text passages mentioning bacteria habitats and species in a corpus of scientific papers abstracts and their normalization to *OntoBiotope Ontology* and to *NCBI Taxonomy*, respectively. Description of the task and the participating systems are available in Del *et al.* (2016).

There are also some works published that try to tackle the challenge of normalization of disease and other clinical concepts. For example, Siu *et al.* (2016) developed a system to perform disambiguation of *Medical Subject Heading* or MeSH terms (used for indexation of PubMed articles) included in the *Unified Medical Language System* (UMLS) KB. The system applies one or more heuristics to deal with ambiguous cases, such as considering as the same entity two different text mentions that are expressed in the singular and the plural forms. Another example is DNorm, a disease name normalization tool (Lu *et al.*, 2013). The systems applies machine learning to learn similarities between text mentions and disease concepts from training data.

Outside of the biomedical domain, *Wikipedia* is one of the most used KB for entity linking, because it contains information about an extensive number of topics. A few examples of works in this subject are Mihalcea & Csomai (2007), Dredze *et al.* (2010), Cucerzan (2011) or Pershina *et al.* (2015).

EL approaches can be either global, local or a mixture of both (Ratinov *et al.*, 2011). Early systems relied on local approaches, where each mention in a document is disambiguated independently by string matching or dictionary look-up

[1] https://www.ncbi.nlm.nih.gov/gene/7941
[2] http://2016.bionlp-st.org/tasks/bb2

algorithms. For example, Bunescu & Pas (2006) proposed a system that disambiguated named entities to *Wikipedia*, first by looking for an exact match for the entity in a dictionary and then using a Support Vector Machines model that compares the lexical context around each entity with a candidate disambiguation *Wikipedia* page.

The global approach, by its turn, aims to disambiguate all mentions in a document simultaneously, presupposing that entities appearing in a document must be somehow related. It is the case of graph-based approaches, that use a graph containing the KB candidates for the mentions and then can rank each node to select the highest scoring one for each mention.

Pershina *et al.* (2015) proposed a graph-based disambiguation model that uses the PPR algorithm. The system leverages both local coherence between an entity and a *Wikipedia* candidate, as well candidate contribution to the global coherence within a document. Each candidate/node in the graph is then ranked and the highest scoring candidate for each entity is selected. The model was evaluated in the dataset AIDA, a gold standard containing *Wikipedia* annotations, and obtained a disambiguation accuracy of 0.9177. Zheng *et al.* (2015) proposed a graph-based method that performs collective EL (i.e. follows a global approach), but also using local features, like the entropy of each node. The model is similar to the method proposed in the present work, in the sense that it is a graph-based method and explore the structure of an ontology to add the egdes of the candidates graph, but our method goes even further by considering the information content of each concept, as well the semantic similarity between nodes to calculate node coherence.

More recent approaches to EL resort to machine learning techiques. For example, the system proposed by Li *et al.* (2017) uses a convolutional neural network (CNN) to rank candidates for entities in biomedical articles and clinical records. The system links the entities to medical concepts and obtained a disambiguation accuracy of 0.9030 for SNOMED-CT concepts[1], and 0.8610 for MeSH[2] and OMIM concepts[3].

[1] `https://browser.ihtsdotools.org/?`
[2] `https://www.ncbi.nlm.nih.gov/mesh`
[3] `https://www.omim.org/`

Karadeniz & Özgür (2019) presented an unsupervised approach to EL using word embeddings, semantic similarity and two KB: Onto-Biotope ontology and the Medical Dictionary for Regulatory Activities (MeDRA). Word embeddings are vector representations of a named entity and usually contain the words of the named entity and their context words, i.e the surrounding words (after and before). An entity vector is compared with ontology concept vectors through semantic similarity, and then the concepts are ranked based on the results.

Most of the existant EL systems were developed for general KBs, like *Wikipedia* and the fewer systems developed for the biomedical domain usually are only adapted to a given specific domain, like diseases or gene names. So, it is essential to develop a flexible system, able to perform EL in biomedical ontologies independently of the specific domain. The system must be easily adapted to any domain in the biomedical sciences, since many biomedical ontologies have a standard structure. Besides this, it is an advantage if the system can also be adapted to more general domains, because it allows a better comparison of its performance with other state of the art systems. Additionally, there is not many training data available for the biomedical domain, so one way of overcoming this obstacle is to use PPR-based methods, that do not require training data. The system proposed by Pershina *et al.* (2015) is PPR-based and uses the context of the entities to disambiguate them, but it is not adapted to biomedical domain and it does not use SSMs or ontologies to build the candidates graph. Instead, the method uses the hyperlink structure of *Wikipedia*, which can not be replicated in specific biomedical domains due to the absence of equivalent KBs. Due to the large availability of ontologies in the biomedical domain, these constitute a better option to help to build the candidates graph. Zheng *et al.* (2015) proposed a system that uses the structure of ontologies to build the graph, but uses the concept of entropy of relations, instead of PPR or SSMs. So, a system able to combine these features can potentially improve the EL task. The valuable information present in biomedical ontologies can be leveraged for the following aims: i) to build the candidates graph for PPR application and ii) to allow the application of SSMs, which can determine the candidates that fit better in the graph.

Chapter 3

Methods

3.1 Problem definition

A knowledge base (KB) is defined as a tuple $< C, R >$, where C is the set of concepts and R the set of relationships between concepts.

The entity linking (EL) task can be divided in two distinct steps:

1. Generation of KB candidate list for each entity e in E (set of named entities, for example, in a document):

$$CL(e) = \{{c_e}^1, ..., c_e^i | c_e \in C\}$$

2. Selection of the KB candidate $c_e \in CL(e)$ that best represents each e.

3.2 Graph building exploring ontology structure

After the step(1) of Section 3.1, a graph G is created:

$$G = \big\{(e, c_e) | e \in E, c_e \in CL(e)\}$$

The graph nodes (e, c_e) are pairs entity/candidate. The graph edges are added based on the structure of the ontology, more specifically, according to the length of paths between two given concepts.

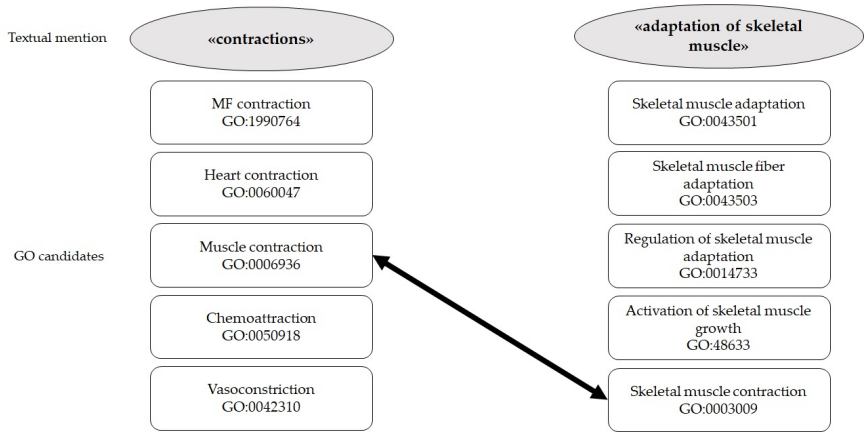


Figure 3.1: Adding an edge in the graph, represented by the black double arrow. Left: mention "contraction" and respective candidate list. Right: mention "adaptation of skeletal muscle" and respective candidate list. The ontology distance between the GO terms "Muscle contraction" and "Skeletal muscle contraction" is two. The maximum distance allowed (a parameter that can be changed) is two, so an edge is built between the nodes that correspond to the referred concepts

An edge in the graph between two nodes occurs if the ontology concepts represented by the nodes are within a maximum distance allowed (in the ontology). Besides that, a link can only occur between two candidates of different entities, because only one element in the candidate list can be the chosen one (Figure 3.1). For each document in corpus there will exist a distinct set of named entities (E) and a distinct graph (G).

3.3 Graph-based disambiguation with Personalized PageRank (PPR)

In step(2) of Section 3.2, the function *disambiguate* selects the correct candidate c_e for entity e :

$$disambiguate(e) = arg_{c_e} max\{score(e, c_e)\} \tag{3.1}$$

The *score* function above determines the likelihood of a given candidate being

the correct disambiguation for given entity e.

Following the approach described in Pershina *et al.* (2015), it is possible to quantify the global coherence of each node in the graph G previously built. For example, if a given node (e, c_e) presents a high coherence to the graph, it is likely that the candidate concept represented by c_e constitutes the correct disambiguation for the respective entity e. The PPR algorithm simulates random walks across graph G with a certain damping factor (see Section 2.4) and ranks the importance of each node.

The calculation of the global coherence of a node n to the graph comprises two steps. First, PPR algorithm determines the coherence of node n to each source node s:

$$coherence_s(n) = PPR(s \rightarrow n) \tag{3.2}$$

Second, the sum of the coherence of node n to each source node s returns the global coherence for node n:

$$coherence(n) = \sum_{s \in G} coherence_s(n) \tag{3.3}$$

Thus, for each entity e, the application of PPR method will chose the node (e, c_e) that presents higher coherence/rank among the competing candidates/nodes, which will correspond to the *score* function (3.1).

3.4 Information Content

According to Resnik (1995), it is possible to associate probability with concepts in an ontology to measure the specificity of each concept.

An ontology C can be viewed as a function $p : C \rightarrow [0, 1]$, where for each $c \in C$, there is a p(c) corresponding to the probability of encountering an instance of the referred concept. Consequently, the Resnik's expression to quantify the *information content* of a concept is given by:

$$IC(c) = -log(p(c)) \tag{3.4}$$

Intuitively, information content can be viewed as measure of "rareness" : as the probability of finding an instance of a given concept rises, its information content declines. An ontology contains subsumption relationships, so, to find a concept is equivalent to implicitly find all of its ancestors. Hence, the probability of finding an instance of the root concept is 1, because this concept is an ancestor of every concept in taxonomy, and its information content is 0.

The probability function p can be `intrinsic`, where p corresponds to the number of its child nodes or `extrinsic`, being p defined according to the frequency of each concept in an external dataset (Couto & Lamurias, 2018b). In the present work, p is defined in an extrinsic manner.

Considering an ontology or KB represented by $< C, R >$ (C is the set of concepts, R the set of relationships), an external dataset by D and a predicate $refer(d, c)$ that is true when an element $d \in D$ refers the concept $c \subset C$, the frequency of concept c is:

$$F_D(c) = |d : refer(c_1, d) \wedge d \in D \wedge c_1 \in Desc(c) \cup c|$$

With the frequency measure F_D, it is possible to calculate the `extrinsic probability` for each concept in the ontology:

$$p(c) = \frac{F_D(c) + 1}{max\{F_D(c_1) : c \in C\} + 1} \tag{3.5}$$

Including equation (3.4) (local coherence) in the coherence equation (3.2) renders:

$$coherence_s(n) = PPR(s \rightarrow n) \cdot IC(n) \tag{3.6}$$

3.5 Semantic similarity

Semantic similarity measures (SSMs) determine the similarity of two distinct concepts in a semantic base (like an ontology). SSMs are restricted to subsumption relationships (*is-a*), which are transitive: for example, assuming that R is the set of relationships in the ontology and knowing that $(c_1, c_2) \in R$ and $(c_2, c_3) \in R$, then we can assume that $(c_1, c_3) \in R$. The ancestors of a concept c are thus defined by:

$$Anc(c) = \{a : (c : a) \in T\}$$

Being T the smallest relation in C (set of concepts) that contains R and are transitive. The common ancestors (CA) of two concepts are defined as:

$$CA(c_1, c_2) = Anc(c_1) \cap Anc(c2)$$

According to Couto & Lamurias (2018b), instead of using the totality of the common ancestors between two concepts, usually SSMs resort to either :

1. the most informative common ancestors ($MICA$):

$$MICA(c_1, c_2) = \{a : a \in CA(c_1, c_2) \wedge IC(a) = \max\{IC(a) : a \in CA(c_1, c_2)\}\}$$

2. or the disjunctive common ancestors (DCA), if one considers multiple inheritance relationships:

$$DCA(c_1, c_2) = \{a : a \in CA(c_1, c_2) \wedge \forall_{a_x \in CA(c_1, c_2)} PD(c_1, c_2, a) = PD(c_1, c_2, a_x) \Rightarrow IC(a) > IC(a_x)\}$$

   being PD a function that determines the difference between the number of paths of c_1 and c_2 to their respective CA.

SSMs can be defined in terms of the IC of the concepts. Resnik (1995) first proposed a SSM based on the idea of shared IC between two concepts, that corresponds to the average of IC of either the MICA or the DCA:

$$SSM_{resnik}(c_1, c_2) = IC_{shared}(c_1, c_2) \tag{3.7}$$

Lin (1998) proposed the folowing SSM:

$$SSM_{lin}(c_1, c_2) = \frac{2 \times IC_{shared}(c_1, c_2)}{IC(c_1) + IC(c_2)} \tag{3.8}$$

Additionally, Jiang & Conrath (1997) proposed a SSM based on the distance between two concepts in an ontology:

$$dist_{jc}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC_{shared}(c_1, c_2)$$

The idea behind it is that the distance between two concepts is inversely proportional to their similarity, so:

$$SSM_{jc}(c_1, c_2) = \begin{cases} \frac{1}{dist(c1,c2)}, \text{if } dist > 0 \\ 1, \text{otherwise} \end{cases} \tag{3.9}$$

Semantic similarity (local coherence) can thus improve the calculation of a node coherence to a given source node:

$$coherence_s(n) = PPR(s \to n) \cdot IC(n) \cdot SSM(s, n) \tag{3.10}$$

3.6 Models

The approach of the present work is to apply four different models to determine which one achieves higher EL accuracy:

1. `Top match`: the baseline approach. For each textual mention in a document, this model selects the ontology concept that is more similar based on Levenshtein distance, without considering semantics or context. Candidates below a predefined minimum matching score (parameter *min_match_score*) were excluded.

2. `PPR`: corresponds to the application of PPR algorithm to EL as described in Pershina *et al.* (2015), but adapted to biomedical ontologies. The coherence of each node to a given source node is calculated by equation (3.2).

3. `PPR-IC`: model that applies the PPR algorithm to EL but includes the information content of each node to calculate the coherence (3.6). The parameter *maximum distance* between ontology concepts that is used in graph building (see Section 3.2) was optimized through the application of PPR-IC model with different maximum distance values.

4. `PPR-SSM`: model that includes the calculation of semantic similarity between two given nodes to obtain node coherence (3.10). The SSM considered are Resnik's (3.7), Lin's (3.8) and Jiang and Conrath's (3.9). Each SSM is calculated using the MICA or the DCA.

3.7 Evaluation

The disambiguation accuracy obtained by a model in a dataset is calculated through the following equation:

$$tp : True\ positives\ or\ number\ of\ entities\ correctly\ disambiguated$$
$$total\_disambiguated : Total\ number\ of\ entities\ disambiguated$$

$$accuracy = \frac{tp}{total\_disambiguated} \tag{3.11}$$

The recall obtained by a model in a dataset is obtained by the expression:

$$fn : False\ negatives\ or\ number\ of\ entities\ not\ disambiguated$$

$$recall = \frac{tp}{tp + fn} \tag{3.12}$$

The models were integrated in a unique system developed with Python programming language, version 3.6.8, where it is possible to choose which model to apply. The next step is to evaluate the system and each of the models in a dataset containing annotations, i.e., the surface form of the entities to disambiguate and the respective KB disambiguations. A disambiguation contains the most appropriate KB concept identifier for the respective entity. The main question at this point is: which model will achieve higher disambiguation accuracy?

Chapter 4

Results and Discussion

4.1 Data

The models previously described (Section 3.6) were evaluated in two different gold standards: CRAFT and MSNBC. These gold standards are corpus containing documents manually annotated with ontology concepts.

The "Colorado Richly Annotated Full-Text" (CRAFT) corpus is a collection of 67 biomedical articles from PubMed Central Open Access subset[1] semantically annotated with concepts belonging to several OBOs (Cohen *et al.*, 2017). CRAFT contains, among others, GO and Uberon annotations. The v3.0 release of the corpus was used in the present work and the subsets of the corpus containing GO *Biological Process*, GO *Cellular Component* and Uberon annotations are henceforth designated by CRAFT-BP, CRAFT-CC and CRAFT-UB. GO *Molecular Function* annotations were present in the corpus, but were not considered due to their overall low number and higher repetitive profile.

To evaluate the performance of the models in a general domain outside the biomedical sciences, it was used the MSNBC corpus, which contains twenty news stories spanning ten MSNBC news categories: Business, United States politics, Entertainment, Health, Sports, Tech & Science, Travel, TV news, United States news and World news. Each article contains a variable number of named entities annotated with the Wikipedia articles, for which there is always an equivalent

[1] `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

DBpedia resource. A summary of both CRAFT and MSNBC corpora is shown in Table 4.1.

Table 4.1: Gold standards used for model evaluation. The metric "Entities w/-candidates" is calculated after removal of candidates below minimum matching score (see "Evaluation setup" 4.2)

	CRAFT-BP	**CRAFT-CC**	**CRAFT-UB**	**MSNBC**
Documents	67	67	67	20
Total entities	9280	4075	12269	755
Entities w/ candidates	8822	3849	12263	657
% Entities w/ candidates	0.9506	0.9445	0.9995	0.8702
Entities w/ solution	7173	3436	11135	405
% Entities w/ solution	0.7730	0.8432	0.9076	0.5364
Entities w/solution per document	107.0597	51.2836	166.1940	20.25
Unique entities per document	26.7313	13.2836	38.5224	12.1

Almost all entities in every corpus have at least one candidate in the respective ontology obtained through fuzzy string matching ("% Entities w/ candidates" in Table 4.1). This computational technique finds strings that match a given pattern, even if approximately. In this context, the pattern was the surface form or the entity present in the annotations file. The number of entities with the correct solution in the candidate list is comparatively slower ("% Entities w/ solution" in Table 4.1), which gives margin to improvement in the candidate list generation process, particulary in the case of the entities in MSNBC corpus (only 53,64% of the entities have the correct disambiguation in candidate list).

The metric "Entities w/solution per document" (Table 4.1) is relevant for the application of PPR algorithm, because documents with fewer entities are more prone to disambiguation errors. So, the higher the number of entities in a document, more "context" exists to assist the disambiguation. Although this metric apparently has high values in all corpora considered, the truth is that the number of unique entities in each document ("Unique entities per document" in Table 4.1) is relatively low, reaching its lowest number of 12.1 in MSNBC corpus. This happens due to the great number of repeated entities present in each document.

In this work it was used the February 2019 release of GO, the Version 2018-11-25 of Uberon and the Version 2016-10 of DBpedia.

4.2 Evaluation setup

Between every candidate in the list and its respective entity it was calculated a matching score based on edit distance. Candidates below minimum matching score (represented by the parameter `min_match_score`) were excluded of the candidate list. In CRAFT-BP, CRAFT-CC and CRAFT-UB, `min_match_score` was adjusted to 0.6 and in MSNBC to 0.2 (based on F1 score obtained by Top Match model).

Some candidates were not linked directly to any other candidate according to ontology relationships, which consequently hampers the random walks performed by the PPR algorithm. The parameter `max_dist` (maximum distance allowed) was included in the system and the effect of its variation between 1 and 8 was studied in the PPR-IC model. For example, if the maximum distance allowed between two ontology concepts were 2, there would be an edge between the candidates represented by those concepts only if the shortest path between the two concepts in the ontology was equal or smaller than 2.

PPR was computed according to the Monte Carlo algorithm proposed by (Fogaras & Rácz, 2004). The parameters for PPR were set according to (Pershina *et al.*, 2015): initialization with 2,000 random walks for each source node, 5 steps of PPR and probability of jump to the source node (or teleport probability) of 0.2.

To perform the disambiguation the function 3.1 was applied for each entity. This function selected the candidate with the highest global coherence (equation 3.3), which by its turn was calculated through the equations 3.2, 3.6 and 3.10. For each of the datasets referred in Section 4.1 the models described in the Section 3.6 were applied. If the selected candidate corresponded to the annotation in the gold standard, the disambiguation was considered correct. In the end, the disambiguation accuracy was calculated for each dataset, that is equivalent to the number of correct disambiguated entities divided by the total number of entities.

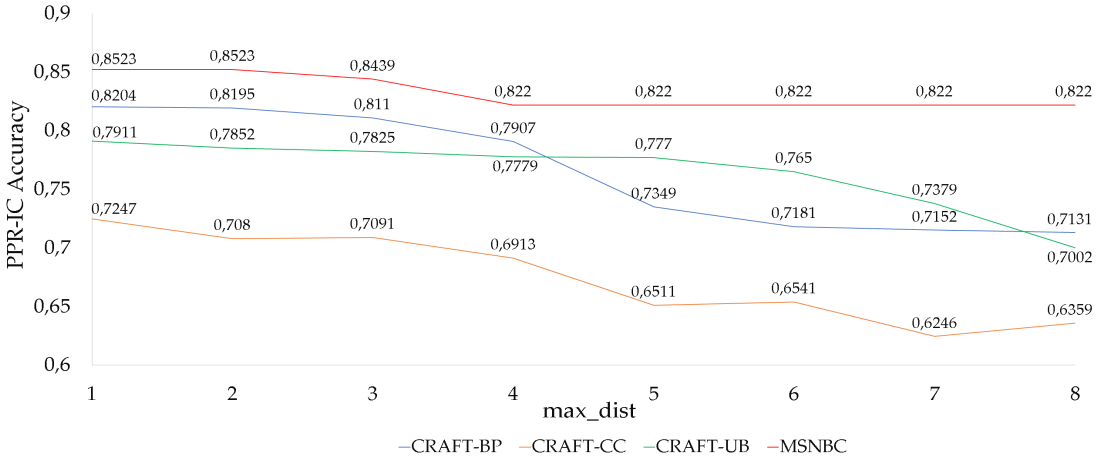


Figure 4.1: Study with PPR-IC model to determine the optimal value for the parameter `max_dist`

In addition, for the model PPR-SSM it was tested the effect of using different SSMs in the disambiguation accuracy, like Resnik's, Lin's, and Jian and Conrath's measures, and the effect of using either the MICA or the DCA.

4.3 Evaluation results

In all datasets, the optimal value for the parameter `max_dist` is 1 (Fig. 4.1). The best results were 0.7911 for CRAFT-BP, 0.7247 for CRAFT-CC, 0.8204 for CRAFT-UB and 0.8523 for MSNBC. The optimal values were then used in all models involving PPR.

The recall obtained by PPR-SSM model for each dataset is 0.6363 for CRAFT-BP, 0.7809 for CRAFT-CC, 0.8316 for CRAFT-UB and 0.4674 for MSNBC. Disambiguation accuracy obtained by the different models on the referred datasets are in Table 4.2.

The model PPR-SSM achieved higher EL accuracy in all datasets, more specifically, 0.8244 in CRAFT-BP, 0.7258 in CRAFT-CC, 0.7918 in CRAFT-UB and 0.8814 in MSNBC.

Results for the PPR-SSM model, using different SSMs and either the MICA or the DCA are in Table 4.3.

Table 4.2: Disambiguation accuracy of each model (rows) in the datasets CRAFT-BP, CRAFT-CC, CRAFT-UB and MSNBC, with the best result for each gold standard highlighted

Model	CRAFT-BP	CRAFT-CC	CRAFT-UB	MSNBC
Top match	0.7744	0.6899	0.7400	0.8471
PPR	0.6926	0.6166	0.7275	0.8432
PPR-IC	0.8204	0.7247	0.7911	0.8523
PPR-SSM	**0.8244**	**0.7258**	**0.7918**	**0.8814**

Table 4.3: Disambiguation accuracy of the PPR-SSM model in each gold standard using different SSMs (Resnik, Lin and JC) and different ways to take into account the IC of the ancestors (MICA or DCA). Best result for each gold standard in bold.

SSM	IC_{shared}	CRAFT-BP	CRAFT-CC	CRAFT-UB	MSNBC
Resnik	MICA	0.7444	0.6439	0.7460	0.8776
	DCA	0.7545	0.6439	0.7471	0.8776
Lin	MICA	**0.8244**	0.7190	**0.7918**	**0.8814**
	DCA	0.8216	**0.7258**	**0.7918**	**0.8814**
JC	MICA	0.8177	0.6985	0.7829	0.8771
	DCA	0.8199	0.6997	0.7899	0.8771

Lin's measure using MICA yielded the best result in CRAFT-BP dataset, Lin's measure using DCA in CRAFT-CC and both Lin using either MICA or DCA achieved the best results in CRAFT-UB and MSNBC datasets.

Additionally, for a better interpretation of the results of the present work, the evaluation results of the models in two datasets containing ChEBI and HPO annotations are shown in Table 4.4.[1]

[1]Evaluation results obtained by André Lamúrias

Table 4.4: Accuracy of the different models on the ChEBI-patents and HPO-GSC gold standards, with the best result for each dataset highlighted

Method	ChEBI-patents	HPO-GSC
Top match	0.5271	0.638
PPR	0.6654	0.5544
PPR-IC	0.8026	0.6557
PPR-SSM	**0.8039**	**0.6825**

4.4 Discussion

Analysing the errors in the disambiguations performed by the PPR-SSM model, one that stands out is the lack of edges in some of the candidates in the graph (particularly in CRAFT datasets), meaning that many nodes/candidates are isolated in the graph. For example, all the candidates for the entity `regulation of blood flow` (in document "11532192" of the CRAFT-BP dataset) were not linked to any other candidates in the document. In cases like this, the PPR algorithm is not applied, instead the candidate with more descendants is chosen. This approach does not take into account the global coherence in candidate selection, which leads to systematic disambiguation errors. This also explains the poor results obtained by the PPR model in all datasets, even when compared with the baseline approach.

Relaxing the parameter `max_dist` in order to capture indirect relationships between concepts in the ontology (i.e. shortest paths between concepts with length greater than 1) and consequently to increase the number of edges in the graph does not seem to be effective, as attested by the results depicted in Fig. 4.1. In fact, increasing `max_dist` value decreases the disambiguation accuracy, which is due to the fact that many of the additional edges are noise, linking candidates that are not the correct choice and thus increasing their overall score.

In all CRAFT datasets there were also issues related with child and parent concepts and their respective IC. There were many cases where two candidates, one being the child of the other in the context of the respective ontology, were "tied": the candidates had the same number of edges connecting to other nodes

and the same contribution for local coherence (determined by the SSM in equation 3.10) and so the PPR-SSM model could not select one candidate. To break the tie between the two candidates, the candidate with higher IC would be chosen, even if it was not the correct disambiguation. For example, in the document "16103912" of the CRAFT-UB dataset, the terms `respiratory organ` (UBERON:0000171) and `respiratory system` (UBERON:0001004) are both candidates for the entity `respiratory`, being `respiratory system` the annotation in the gold standard. Both terms had no links to other candidates in the document but `respiratory organ` was the child term of `respiratory system`, thus had an higher IC. The model wrongly disambiguated the entity with the candidate `respiratory organ` due to the weight of IC. A similar case ocurred with the entity `antibodies` in CRAFT-CC dataset (document "11897010") and the respective candidates `IgG immunoglobulin complex, circulating` (GO:0071736) and `antibody` (GO:004 2571). The term GO:0071736 is a child concept of GO:0042571, and because it had an higher IC, it was selected to the detriment of the correct term. However, in other tied cases the candidate with higher IC is selected, which corresponded to the correct disambiguation. For example, `long-chain fatty acid metabolic process` (GO:0001676) and `long-chain fatty acyl-CoA metabolism` (GO:003 5336) were both candidates for `long chain fat metabolism`, having no links to any other candidate in the graph. The candidate `long-chain fatty acid metabolic process` had higher IC and it was selected, which was the correct disambiguation. Consequently, it is not possible to derive a general rule to solve all cases.

Another source of errors lies on the step of candidates list generation, specially for MSNBC gold standard, considering that the PPR-SSM model achieved a low recall (0.4674) in this dataset. In this way, it was explored the hypothesis that the stemming (i.e., to convert inflected words to their root form) of the surface forms could improve the retrieval of the candidates list. If this approach was successful, more entities would see the correct disambiguation in the respective candidates list and thus the disambiguation accuracy would be higher. In some cases, the impact of these approaches was direct, as it was the case of the entity `vessels` in the CRAFT-UB dataset. Without stemming the surface form, the system selected the wrong candidate `blood vessels` (UBERON:0004537), but

with the stemming of the surface form *vessels* to *vessel*, there was an exact match with the candidate `vessel` (UBERON:0000055). After the implementation of this stemming step in the workflow, the conclusion was that it was not time-effective, in the sense that the running time of system greatly increased and the impact on the disambiguation accuracy was almost negligible.

The low recall achieved by PPR-SSM model in the MSNBC dataset is also explained by the exclusion of some entities from candidates list. It was decided that *DBpedia* candidates that were not instantiated in any *DBpedia* ontology class would be eliminated from the candidates list where they appeared. Otherwise, the application of SSMs would not be possible, because these candidates were not represented in the structure of the ontology. The downside was that many candidates that were the correct disambiguation for the respective entities were excluded and so the entities were not disambiguated. This lack of classification for many *DBpedia* instances is due to the fact that the *DBpedia* ontology is very small and shallow: the *DBpedia* KB contains more than 4 million instances but the ontology only has 685 classes, so many instances are not classified. This is related with the nature of the project, since it is maintained and developed by voluntary online users, resulting in few contributions, absence of classification for many instances and in classifications errors, problems already pointed by Paulheim & Bizer (2013). For example, the concept `Sweden`[1] is classified as `MusicalArtist` and not as `Country` (as the time of writing).

For comparison with state-of-the-art approaches, mainly developed for social media text and applied to the disambiguation of Wikipedia entities, the system was tested in the MSNBC dataset, a corpus annotated with *Wikipedia* entities. As the system requires an ontology, the *Wikipedia* entities were disambiguated to the *DBpedia* ontology, as both *Wikipedia* and *DBpedia* share the same instances. Recently, Guo & Barbosa (2018) proposed two EL algorithms: `WNED`, an iterative greedy algorithm based on random walks and `L2R.WNED`, a learning-to-rank algorithm trained on benchmark corpora. The authors determined the disambiguation accuracy obtained by the referred approaches on the MSNBC dataset and also by seven state-of-the-art systems. `L2R.WNED` obtained a new state-of-the-art accuracy of 0.91, whereas the state-of-the-art approaches achieved accuracy

[1] `http://dbpedia.org/resource/Sweden`

values ranging between 0.66 and 0.89. Hence, the accuracy obtained by the PPR-SSM model (0.8814) in the present work compares with the accuracy obtained by state-of-the-art systems.

The impact of the SSMs in the three CRAFT datasets was modest as both PPR-IC and PPR-SSM models achieved similar accuracy. In these three cases, the PPR-IC model was able to increase the accuracy by 0.0500 in CRAFT-BP, 0.0348 in CRAFT-CC and 0.0511 in CRAFT-UB comparing to the baseline approach, contrasting with the increase of 0.0052 in MSNBC. By its turn, comparing to the PPR-IC model, the PPR-SSM model increased the accuracy by 0.0040 in CRAFT-BP, 0.0011 in CRAFT-CC, 0.0007 in CRAFT-UB and 0.0291 in MSNBC. This suggests that SSMs application is particularly useful in cases where the PPR-IC model does not perform so well, in this case in the MSNBC dataset. If the accuracy obtained through PPR-IC is already high, the impact of SSMs will be limited. Results obtained in ChEBI and HPO datasets with other modules of the system here proposed corroborate this hypothesis (Table 4.4). For ChEBI dataset, the PPR-IC model achieved an increase of 0.2755 in accuracy relatively to baseline approach, so PPR-SSM accuracy improvement besides PPR-IC was small (0.0013). On the other hand, for HPO dataset, PPR-IC only improved accuracy in 0.0177, allowing an improvement of 0.0268 by PPR-SSM model besides PPR-IC.

The model PPR-SSM achieved higher disambiguation accuracy in all datasets when compared to the other models evaluated (see Table 4.3), thus it is possible to conclude that the use of SSMs can have a positive impact in EL task. This is specially valid to the MSNBC dataset, where the model improved disambiguation accuracy by 0.0291 comparing to the second best performing model, PPR-IC.

Lin's measure demonstrated to be the best SSM (see Table 4.3), in opposition to Resnik's. The main difference with Resnik's measure is that it does not consider the IC of individual concepts, which may lower the accuracy results.

Nevertheless the differential impact of SSMs in distincts datasets, another conclusion extracted from the results is that it is possible to explore the semantic relationships encoded in an ontology to generate the candidates graph and perform EL as a ranking task. This is true for EL in biomedical field, specifically for

the disambiguation of gene functions and anatomical parts, but also for a general domain, like the disambiguation of *DBpedia* entities.

Although not the focus of the experiments here described, it is important to note the high recall achieved by the system in CRAFT datasets (0.6363 for CRAFT-BP, 0.7809 for CRAFT-CC, 0.8316 for CRAFT-UB), while simultaneously obtaining a good disambiguation accuracy. For comparison, another system proposed by Boguslav *et al.* (2018) achieved a recall of less than 0.20 in CRAFT-BP and less than 0.70 in CRAFT-CC datasets.

Chapter 5

Conclusions

The PPR-SSM model achieved the highest disambiguation accuracy in all datasets evaluated: 0.8244 in CRAFT-BP, 0.7258 in CRAFT-CC, 0.7918 in CRAFT-UB and 0.8814 in MSNBC. This demonstrates that is possible to improve the performance of EL applying the PPR algorithm, SSMs and using biomedical ontologies.

PPR-SSM model generates a list with ontology candidates for each entity in a document, builds a graph with the candidates as nodes and leverages the structure of biomedical ontologies to generate the edges, in contrast with the PPR-based method proposed by Pershina *et al.* (2015), that leverages the hyperlink structure of *Wikipedia*. The model then determines the coherence of each node in the graph, i.e., how well the node fits into the graph. The coherence of each node/candidates is calculated through the application of the PPR algorithm, the IC of the candidate and the application of an SSM to determine the semantic similarity between the candidate and every one of the candidates for other entities. After this, the candidates are ranked and for each entity the highest ranking candidate is selected as the correct disambiguation.

PPR-SSM combines the advantages of PPR-based methods without requiring training data, which confers adaptability to other KB, as well other domains beyond biomedical text mining. In fact, the model performed well in the disambiguation of entities to a general ontology, *DBpedia*, obtaining an acuracy of 0.8814 on MSNBC dataset. Most state of the art systems obtained accuracy values ranging from 0.66 to 0.91 on the same dataset (Guo & Barbosa, 2018),

which demonstrates that PPR-SSM performance is comparable to those obtained by state of the art systems.

The achieved results also demonstrate that it is possible to explore the structure of biomedical ontologies to build the candidates graph in PPR-based methods. However, there are some shortcomings, as many graphs that were built contained many isolated nodes, which hindered the performance of the PPR algorithm. This aspect of the system needs to be improved in future modifications.

The PPR-SSM model achieved a high recall in CRAFT datasets: 0.6363 in CRAFT-BP, 0.7809 in CRAFT-CC, 0.8316 in CRAFT-UB. On the other side, the recall obtained by the model in the MSNBC dataset was low: 0.4674.

This work led to the development of a biomedical EL module and the software is available at `https://github.com/lasigeBioTM/PPRSSM`.

5.1 Future work

The problem of the lack of edges in the candidates graph can be approached by extracting relationships between concepts described in the literature that are not represented in ontology's structure. The approach proposed by Lamurias *et al.* (2019) to detect and classify relationships described in biomedical text can be adapted to GO and Uberon entities. If this strategy works, there will be less nodes isolated in the candidates graph and the random walks in the PPR algorithm will be possible in more parts of the graph. So PPR application will be more effective, improving the disambiguation accuracy.

Specially for the entities in the MSNBC dataset where the module achieved a low recall, another problem associated with the module is the absence of the correct candidate in the candidates list for some entities. The module relies on string matching to generate the candidates list from the ontology for each entity, but Prokhorov *et al.* (2019) proposed an approach in which entities are represented as graph paths containing all of their ancestors in the ontology and the representation is then used to retrieved the candidates list. This sytem converts ontologies from directed acyclic graphs into rooted tree graphs and then it uses deep learning, more concretely Long-Short-Term Memory (LSTM) networks to represent the entities as vectors and to compare them with the rooted tree

graph. This strategy can possible improve the recall and increase the number of exact matches, even in CRAFT datasets, which will ultimately increase the disambiguation accuracy in all datasets.

EL is closely associated with the Named Entity Recognition (NER) task, as it necessarily requires the identification of the entities in the text before their linking to a KB. In this way, it would be useful to combine the developed module with a NER system in order to create a tool capable of doing both EL and NER, like for example the system MER proposed by Couto & Lamurias (2018a).

References

Agirre, E. & Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings ofthe 12th Conference ofthe European Chapter ofthe ACL*, April, 33–41, Association for Computational Linguistics. 16, 17

Alonso, I. & Contreras, D. (2016). Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents : An UMLS approach. *Expert Systems With Applications*, **44**, 386–399. 15

Arighi, C., Hirschman, L., Lemberger, T., Bayer, S., Liechti, R., Comeau, D. & Wu, C. (2017). Bio - ID Track Overview. 17, 18

Arp, R., Smith, B. & Spear, A.D. (2015). *Building ontologies with basic formal ontology*. MIT Press. 3, 10

Basaldella, M., Furrer, L., Tasso, C. & Rinaldi, F. (2017). Entity recognition in the biomedical domain using a hybrid approach. *Journal of biomedical semantics*, **8**, 51. 3

Blanco, R., Ottaviano, G. & Meij, E. (2015). Fast and Space-Efficient Entity Linking in Queries Categories and Subject Descriptors. In *WSDM 2015*, 179–188. 7

Boguslav, M., Cohen, K.B., Hunter, L.E. & Program, C.B. (2018). Improving precision in concept normalization. In *Pacific Symposium on Biocomputing.*, 566–577. 38

Bunescu, R. & Pas, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for*, April, 9–16, Trento, Italy,. 19

REFERENCES

Clark, K. & Manning, C.D. (2016). Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings ofthe 2016 Conference on Empirical Methods in Natural Language Processing*, 2256–2262, Association for Computational Linguistics, Austin, Texas. 9

Cohen, K.B. & Hunter, L. (2007). Natural Language Processing and Systems Biology. *Artificial Intelligence Methods And Tools For Systems Biology*, 147–173. 2

Cohen, K.B., Verspoor, K., Funk, C., Bada, M., Palmer, M. & Hunter, L.E. (2017). The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation in the Biomedical Domain The Colorado Richly Annotated Full Text (CRAFT) Corpus : Multi-Model Annotation In The Biomedical Domain. In *The Handbook of Linguistic Annotation*, June. 29

Courtout, M., Mungall, C., Brinkman, R.R. & Ruttenberg, A. (2011). Building the OBO FOundry - One Policy at a time. In *International Conference on Biomedical Ontologies*. 10

Couto, F. (2019). Data and text processing for health and life sciences. In *Advances in Experimental Medicine and Biology*, vol. 1137, Springer, Cham. 1

Couto, F.M. & Lamurias, A. (2018a). MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, **10**, 58. 41

Couto, F.M. & Lamurias, A. (2018b). Semantic Similarity Definition. *Reference Module in Life Sciences*, 0–16. 4, 14, 24, 25

Cucerzan, S. (2011). Tac entity linking by performing full-document entity extraction and disambiguation. In *Proceedings of the Text Analysis Conference*. 18

Daub, J., Gardner, P.P., Tate, J., Ramsko, D., Manske, M., Scott, W.G., Weinberg, Z., Griffiths-Jones, S. & Bateman, A. (2008). The RNA WikiProject : Community annotation of RNA families. *RNA*, 2462–2464. 13

Del, L., Bossy, R., Chaix, E., Ba, M., Ferr, A., Bessi, P. & Claire, N. (2016). Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In *Proceedings ofthe 4th BioNLP Shared Task Worksho*, 12–22. 18

Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J. & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, **51**, 32–49. 3

Dredze, M., Mcnamee, P., Rao, D., Gerber, A. & Finin, T. (2010). Entity Disambiguation for Knowledge Base Population. In *23rd International Conference on Computational Linguistics*, August, 277–285. 7, 18

Ernst, P., Siu, A., Milchevski, D., Hoffart, J. & Weikum, G. (2016). DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences. *Proceedings of ACL-2016 System Demonstrations*, 19–24. 8

Fleuren, W.W. & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, **74**, 97–106. 2

Fogaras, D. & Rácz, B. (2004). Towards Scaling Fully Personalized PageRank. In L. S., ed., *Algorithms and Models for the Web-Graph*, vol. 3243, Springer,, Berlin, Heidelberg. 31

Gattani, A., Lamba, D.S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V. & Doan, A. (2013). Entity Extraction , Linking , Classification , and Tagging for Social Media : a Wikipedia-Based Approach. *Proceedings ofthe VLDB Endowment*, **6**, 1126–1137. 7

Good, B.M., Clarke, E.L., Alfaro, L.D. & Su, A.I. (2012). The Gene Wiki in 2011 : community intelligence applied to human gene annotation. *Nucleic acids research*, **40**, 1255–1261. 13

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**, 199–220. 3, 10

Guo, Z. & Barbosa, D. (2018). Robust named entity disambiguation with random walks. *Semantic Web*, **9**, 459–479. 36, 39

Haendel, M.A., Balhoff, J.P., Bastian, F.B., Blackburn, D.C., Blake, J.A., Bradford, Y., Comte, A., Dahdul, W.M., Dececchi, T.A., Druzinsky, R.E., Hayamizu, T.F., Ibrahim, N., Lewis, S.E., Mabee, P.M., Niknejad, A. & Ontologies, G. (2014). Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics 2014* , **5**, 1–13. 13

Hastings, J., Matos, P.D., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M. & Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry : enhancements for 2013. *Nucleic acids research*, **41**, 456–463. 14

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. & Steinbeck, C. (2016). ChEBI in 2016 : Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, **44**, 1214–1219. 13

He, J., de Rijke, M., Sevenster, M., van Ommering, R. & Qian, Y. (2011). Automatic Link Generation with Wikipedia: A Case Study in Annotating Radiology Reports. In B. Berendt, A. de Vries, W. Fan, U. Craig Macdonald University of Glasgow, U. Iadh Ounis University of Glasgow & U. Ian Ruthven University of Strathclyde, eds., *CIKM'11: proceedings of the 2011 ACM International Conference on Information and Knowledge Management: October 24-28, 2011, Glasgow, Scotland, UK*, 1867–1876, Glasgow, Scotland, UK. 8

Hirschman, L., Yeh, A., Blaschke, C., Valencia, A., Hirschman, E.L., Yeh, A., Blaschke, C. & Valencia, A. (2005). Overview of BioCreAtIvE : critical assessment of information extraction for biology. *BMC Bioinformatics*, **10**, 1–10. 17, 18

Hliaoutakis, A., Varelas, G., Voutsakis, E. & Petrakis, E.G.M. (2006). Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 1–29. 15

Hu, W., Qiu, H., Huang, J. & Dumontier, M. (2017). BioSearch : a semantic search engine for. *Database*, 1–13. 8

Huang, M., Liu, J. & Zhu, X. (2011). GeneTUKit : a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033. 18

Hunter, L. & Cohen, K.B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular cell*, **21**, 589–94. 3

Jiang, J.J. & Conrath, D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, 19–33, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan. 26

Jovanovi, J. & Bagheri, E. (2017). Semantic annotation in biomedicine : the current landscape. *Journal of biomedical semantics*, **8**, 1–18. 7, 8

Kang, N., Afzal, Z., Singh, B., van Mulligen, E.M. & Kors, J.A. (2012). Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics*, **45**, 423–428. 8

Karadeniz, I. & Özgür, A. (2019). Linking entities through an ontology using word embeddings and syntactic. *BMC Bioinformatics*, **20**, 1–12. 19

Kilicoglu, H. (2018). Biomedical text mining for research rigor and integrity : tasks , challenges , directions. *Briefings in Bioinformatics*, **19**, 1400–1414. 2

Lamurias, A., Sousa, D., Clarke, L.A. & Couto, F.M. (2019). BO-LSTM: Classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, **20**. 2, 40

REFERENCES

Leaman, R., Khare, R. & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal Of Biomedical Informatics*, **57**, 28–37. 8

Lee, K., Shin, W., Kim, B., Lee, S., Choi, Y., Kim, S., Jeon, M., Tan, A.C. & Kang, J. (2016a). HiPub: Translating PubMed and PMC texts to networks for knowledge discovery. *Bioinformatics*, **32**, 2886–2888. 8

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.C. & Kang, J. (2016b). BEST: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS ONE*, **11**, 1–16. 8

Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B. & Huang, D. (2017). CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics*, **18**. 19

Li, L., Liu, S., Li, L., Fan, W., Huang, D. & Zhou, H. (2013). A Multistage Gene Normalization System Integrating Multiple Effective Methods. *Plos One*, **8**, 1–9. 18

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304. 26

Lu, Z., Kao, H.y., Wei, C.h., Huang, M., Liu, J., Kuo, C.j., Hsu, C.n., Tsai, R.T.h., Dai, H.j., Okazaki, N., Cho, H.c., Gerner, M. & Solt, I. (2011). The gene normalization task in BioCreative III. *BMC Bioinformatics*, **12**. 18

Lu, Z., Leaman, R. & Dog, R.I. (2013). DNorm : disease name normalization with pairwise learning to rank. *Bioinformatics*, **29**, 2909–2917. 18

Mazaitis, K., Wang, R.C., Lin, F., Dalvi, B., Bauer, J. & Cohen, W.W. (2014). A Tale of Two Entity Linking and Discovery Systems. 17

Mihalcea, R. & Csomai, A. (2007). Wikify ! Linking Documents to Encyclopedic Knowledge. In *Cikm - Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 233–241, ACM New York, NY, USA ©2007, Lisbon, Portugal. 18

Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.h., Torres, R., Krauthammer, M., Lau, W.W., Liu, H., Hsu, C.n., Schuemie, M., Cohen, K.B. & Hirschman, L. (2008). Open Access Overview of BioCreative II gene normalization isolated cDNA clones encoding the human homologue Humly9 . Text evidence. *Genome Biology*, **9**. 18

Müller, H.M., Van Auken, K.M., Li, Y. & Sternberg, P.W. (2018). Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinformatics*, **19**, 94. 2

Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E. & Haendel, M.A. (2012). Uberon , an integrative multi-species anatomy ontology. *Genome Biology*, **13**, 1–20. 12

Munkhdalai, T., Li, M., Batsuren, K., Park, H., Choi, N. & Ryu, K. (2015). Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *Journal of Cheminformatics*, **7**, S9. 3

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM COMPUTING SURVEYS*, **41**, 1–69. 9

Nunes, T., Campos, D., Matos, S. & Oliveira, J.L. (2013). BeCAS: Biomedical concept recognition services and visualization. *Bioinformatics*, **29**, 1915–1916. 8

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford InfoLab. 15, 16

Paulheim, H. & Bizer, C. (2013). Type Inference on Noisy RDF Data. In *The Semantic Web – ISWC 2013*, 510–525, Springer, Berlin, Heidelberg, Sydney, NSW, Australia. 36

Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y.M., Haug, K., Carroll, A.J., Spalding, D., Paschall, J., Wang, M., Del-Toro, N., Ternent, T., Zhang, P., Buso, N., Bandeira, N., Deutsch, E.W., Campbell, D.S., Beavis, R.C., Salek, R.M., Sarkans, U., Petryszak, R., Keays, M., Fahy, E., Sud, M., Subramaniam, S., Barbera, A., Jiménez, R.C., Nesvizhskii, A.I., Sansone, S.A., Steinbeck, C., Lopez, R., Vizcaíno, J.A., Ping, P. & Hermjakob, H. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology*, **35**, 406–409. 7

Pershina, M., He, Y. & Grishman, R. (2015). Personalized Page Rank for Named Entity Disambiguation. In *Human Language Technologies: The 2015 Annual Conference ofthe North American Chapter ofthe ACL*, Section 4, 238–243. 3, 4, 17, 18, 19, 20, 23, 27, 31, 39

Prokhorov, V., Pilehvar, M.T. & Collier, N. (2019). Generating Knowledge Graph Paths from Textual Definitions using Sequence-to-Sequence Models. 40

Rak, R., Batista-navarro, R.T., Rowley, A., Carter, J. & Ananiadou, S. (2014). Text-mining-assisted biocuration workflows in Argo. *Database*, 1–14. 7

Rao, D., McNamee, P. & Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. In P.J. Poibeau T., Saggion H., ed., *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing.*, 93–115, Springer, Berlin, Heidelberg. 8

Ratinov, L., Roth, D., Downey, D. & Anderson, M. (2011). Local and Global Algorithms for Disambiguation to Wikipedia. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies, 1375–1384, Association for Computational Linguistics Stroudsburg, PA, USA ©2011, Portland, Oregon — June 19 - 24, 2011. 18

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of IJCAI-95*. 23, 26

Sebastian, K., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.p., Gargano, M., Harris, N.L., Matentzoglu, N., Mcmurry, J.A., Osumi-sutherland, D., Cipriani, V., Balhoff, J.P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Beltran, S., Freeman, A.F., Sergouniotis, P.I., Durkin, D., Storm, A.L., Hanauer, M., Brudno, M., Bello, S.M., Sincan, M., Rageth, K., Wheeler, M.T., Oegema, R., Lourghi, H., Rocca, M.G.D., Thompson, R., Castellanos, F., Priest, J., Cunningham-rundles, C., Hegde, A., Lovering, R.C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X.A., David, G., Boerkoel, C.F., Smith, C., Milner, J.D., Klion, A., Carter, M.C., Groza, T., Smedley, D., Haendel, A., Mungall, C. & Robinson, P.N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, **47**, 1018–1027. 14

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, M. & Heilman, J. (2017). Evolution of Wikipedia ' s medical content : past , present and future. *J Epidemiol Community Health 2017;71:1122–1129.*, 1122–1129. 13

Shastry, B.S. (2009). SNPs : Impact on Gene Function and Phenotype. *Single Nucleotide Polymorphisms, Methods in Molecular Biology 578*. 14

Shen, W., Wang, J., Luo, P. & Wang, M. (2013). Linking named entities in Tweets with knowledge base via user interest modeling. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, 68. 3

Shen, W., Wang, J. & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, **27**, 443–460. 7

Singhal, A., Simmons, M. & Lu, Z. (2016). Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Comput Biol 12(11):*, **12**, 1–19. 2

Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M. & Blike, G. (2016). Allocation of Physician Time in Ambulatory Practice : A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, **165**, 753–760. 8

Siu, A., Ernst, P. & Weikum, G. (2016). Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 72–76. 18

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A. & Christopher, J. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integratio. *Nature Biotechnology*, **25**, 1251. 10

Sullivan, R., Leaman, R. & Gonzalez, G. (2011). The DIEGO Lab Graph Based Gene Normalization System. In *10th International Conference on Machine Learning and Applications*. 18

The Gene Ontology Consortium (2019). The Gene Ontology Resource : 20 years and still GOing strong. *Nucleic Acids Research*, **47**, 330–338. 11

The Gene Ontology Consortium, Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25–29. 11

Uschold, M. & Gruninger, M. (1996). Ontologies : Principles , Methods and Applications. *Knowledge Engineering Review*, **11**, 93–136. 10

Wei, C.H., Kao, H.Y. & Lu, Z. (2015). GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Research International*, **2015**. 18

Winkler, W.E. (1999). The State of Record Linkage and Current Research Problems. *US Census Bureau*, 1–15. 9

Zheng, J.G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J. & Ji, H. (2015). Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, **15**, 1–9. 8, 9, 19, 20